" Development of Deep Learning-based Semantic Analysis Algorithms for Social Networks in Big Data Environment."

Proposal submitted for registration of Ph.D. in Computer Science at Faculty of Science & Technology, Swami Ramanand Teerth Marathwada University, Nanded

Submitted By

Mr. Mohammed Ali Mohammed Ali
Aldobai2020@gmail.com

Under the Guidance of

DR . S. N. LOKHANDE

Assistant Professor
School of Computational Science
SRTM University, Nanded, MS, India.

July-2019

Table of Contents

Sr. No	Contents	Page No
1	Introduction	**3**
2	Literature review	**5**
4	Objectives of the study	**8**
5	Research Methodology	**8**
6	Significance of study	**11**
7	Proposed work plan	**11**
8	References:	**12**

Introduction

The advent of online social networks has been one of the most exciting events in this decade. Many popular online social networks such as Twitter, LinkedIn, and Facebook have become increasingly popular. In addition, a number of multimedia networks such as Flickr have also seen an increasing level of popularity in recent years. Many such social networks are extremely rich in content, and they typically contain a tremendous amount of content and linkage data which can be leveraged for analysis, and the richness of this network provides unprecedented opportunities for data analytics in the context of social networks [1].

Therefore, big data refer to a massive amount of data sets collected from the verities of data source that can't be processed using traditional computing techniques. Big Data is not merely a data rather it has become a complete subject which involves various tools, techniques, and framework. The need for big data generated from large companies like Facebook, Yahoo, Google, YouTube, etc. It is a term that refers to dataset whose volume (size), complexity and rate of growth (velocity) make them too difficult to captured, managed, processed or analyzed by conventional technology and tools such as relational databases [2].

Moreover, the term used to refer a high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. [3]

Big data environment is used to acquire, organize and analyze the various types of data. Data that is so large in volume, so diverse in a variety or moving with such velocity is called "big data". [4] So, big data analytics is important that helps organizations harness their data and use it to identify new opportunities. That, in turn, leads to smarter business moves, more efficient operations, higher profits, and happier customers.

In previous studies it has been proven that analysis of big data is gaining several features such as cost reduction for big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data – plus they can identify more efficient ways of doing business thus faster, better decision making. With the speed of Hadoop and in-memory analytics, combined with the ability to analyze new sources of data, businesses are able to analyze information immediately – and make decisions based on what they've learned and the analysis has considered benefit for the user were provided new products and services. With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want.

Davenport points out that with big data analytics, more companies are creating new products to meet customers' needs. [5]

Nowadays, Sentiment Analysis (SA) or Opinion Mining (OM) is the computational study of people's opinions, attitudes and emotions toward an entity. The entity can represent individuals, events or topics.[6]

The rapid increase in data on social media creates a need for mining such data to get valuable insights. The data type can be unstructured with large volumes. Sentiment analysis addresses such need by detecting opinions or emotions on the social media text.

Sentiment analysis can be performed in various domains such as social, medical and industrial applications. Sentiment analysis is an interdisciplinary research field which depends on techniques from Natural Language Processing (NLP), text mining, machine learning, statistics, and information retrieval, the main aim of sentiment analysis or opinion mining is study of people's opinions, behaviors, emotions, attitudes and beliefs about an entity such as product, event/topic, person or organization.[7]

Literature review

In recent years, Big Data analytics has become popular research in computer science and other related data mining areas. It is a very important research area in new trends and also creates enormous vibration in technocrats. It is not only vibrant in computer science but also spread into many other diversified areas, high volumes of valuable data can be easily collected and generated. Social networks are examples of generating sources of these big data. The analysis of the big volume of user-generated content in social media can provide meaningful information for gleaning people's emotions and a deeper understanding of public attitude and mood.

Perikos et al. have presented a generic framework for the analysis of big social data and the modeling of public emotions and mood. The framework consists of two main parts. Initially, an ensemble classifier schema, which combines a generic domain independent, knowledge-based tool with machine learning methods, is used to recognize emotional content in user-generated social data. Then, a graph-based method is utilized to model the emotional level of a topic based on the emotions recognized and then it creates the topic's emotional graph visualizing public emotions and mood on the topic. [8]

QingwuHuhave et al. have proposed to explore the controversial scientific problem that "The sampling large data in time and space is not the more extensive, the better". The concept of efficient selection (ES) of big data is proposed to extract small sample datasets with the perspective of time and space from a large dataset. Through analyzing the mining potentiality of the small sample datasets on specific function model, select out representative small datasets to replace the original data. This method solves the blindness of acquisition of the traditional small sample data to some extent and provides a powerful direction and method support for the mining of human behavior pattern under the background of large data. [9]

VATRAPU et al. have presented a new approach to big data analytics called social set analysis. The social set analysis consists of a generative framework for the philosophies of computational social science, the theory of social data, conceptual and formal models of social data, and an analytical framework for combining big social data sets with organizational and societal data sets. Three empirical studies of big social data are presented to illustrate and demonstrate social set analysis in terms of fuzzy set-theoretical sentiment analysis, crisp set-theoretical interaction analysis, and event studies- oriented set-theoretical visualizations. Implications for big data analytics, current limitations of the set-theoretical approach, and future directions are outlined. [10]

Huber et al. have proposed and addresses the needs of the manufacturing and plant industry in terms of the processing of a huge amount of complex event data from process monitoring. It combines machine learning and semantic technologies, which allows not only the automatic detection of process failures and their root causes; it also renders these findings in a human-interpretable way thanks to the semantification. By means of a real-world use case, where large-scale printing machines are monitored. [11]

Chhabra et al. have presented a new sub-linear solution based on filtration. Any algorithm for exact string matching can be used as a filtering method. If the filtration algorithm is sub-linear, the total method is sub-linear on average. They show by practical experiments that the new solution is more efficient than earlier algorithms [12].

Jiangbo Lu et al. presented a generic and fast computational framework for general multi-labeling problems called PatchMatch Filter (PMF). They explore effective and efficient strategies to weave together these two fundamental techniques developed in isolation, i.e., PatchMatch-based randomized search and efficient edge-aware image filtering. By decomposition an image into compact super-pixels, they also proposed superpixel-based novel search strategies that generalize and improve the original PatchMatchmethod. Further motivated to improve the regularization strength, they proposed a simple yet effective cross-scale consistency constraint, which handles labeling estimation for large low-textured regions more reliably than a single-scale PMF algorithm. Focusing on dense correspondence field estimation in this paper, they demonstrate PMF's applications in stereo and optical flow. Their PMF methods achieve top-tier correspondence accuracy but run much faster than other related competing methods, often giving over 10-100 times speedup. [13]

TasleemNizam et al. have presented the concept of Big Data and various technologies has been surveyed which are used to handle the big data. This paper discussed the architecture of Big Data using Hadoop HDFS distributed data storage under which its different components are also explained. The main objective of this paper was to make a survey of various Big Data architecture, its handling techniques which handle a huge amount of data from different sources and improves the overall performance of systems and its applications which shows its importance and uses in the present IT world. [14]

Ahmed et al. have proposed sentiment analysis levels, techniques, enhancement methods, applications, list of APIs, lexicons, tools and existing research gaps. Future work will

consider comparing the state of the art techniques presented through this paper using the same data set across all the different techniques to be able to evaluate the best techniques used. [7]

Markić et al. have proposed shows one possible integration of information technology and data mining tools, with the goal of visualizing the attitudes and opinions on the social networks in the form of a word cloud, which can then further be used to create marketing strategies and improve customer relations and customer service.. [15]

Maddumage et al. have proposed paper demonstrates the use of R in SNA. Further, SNA techniques and concepts, types of networks have been included in this paper. According to the literature, different R packages can be applied successfully for SNA. There are no any software packages available in R that cover more functions. Therefore all existing packages which can be used for SNA have their own features. Among all the packages, IGraph mainly provides graphics functions, It also handles large complex networks. All of these packages can be used to calculate centrality, clustering coefficient, network diameter, page level, and density. However, if we need to calculate some specific features, we need to choose different software. If we need to calculate Cohesion, we can use the Networkx or IGraph tools. In IGraph, centrality, page ranking, graphics level and group execution time are minimal. IGraph is therefore better package for SNA when considering the execution time.[16]

Jie Hu et al. have proposed a semantic model that can natural represent various academic social networks, especially various complex semantic relationships among social actors. This model can be used as the foundation for managing, manipulating and querying academic social networks. They also introduce a concise language to represent and query academic social networks.[17]

Leenen et al. have presented argues that both big data analytics and semantic technologies are necessary to provide counter measures against cyber threats. An overview of the use of semantic technologies and big data technologies in cyber defence is provided, and important areas for future research in the combined domains are discussed.[18]

All previous literature reviews which mentioned interested in the process to fetch big data from social network these data have been used to analyze and predict in order to benefit from them in their lifecycle. These analyses and predictions are utilized by the help of machine learning algorithms in order to reach the results and the speed which characterized by the studies mentioned above.

Objectives of the study

The main objectives of this research is to get a full understanding of the area of semantic analysis of social networks in big data environment so as to contribute new solutions for the betterment of the field. The important objectives of the current study are given below:

- To obtaining a comprehensive understanding of semantic analysis, deep learning, and social networks along with big data.
- To provide an overview of the deep learning-based semantic analysis for social networks in term of big data on the basis of some recent related research works.
- To explore various analysis approaches of analyzing social networks data in big data environment and extracting the main challenges of the field of study.
- To present optimal solutions for different extracted challenges of data analysis in big data environment.
- To design and implement new solutions for semantically analyzing big data with help of deep learning.
- To evaluate and analyze the performance of the developed solutions in order to verify its efficiency and effectiveness with comparing them to some recent related solutions have been done in the field of the current study.

Research Methodology

- Study of the different existing techniques and algorithms that are related to semantic analysis for big data in social network
- Developing a general analysis framework, for social network in big data environment .
- Using any simulator / analyzing tools in order to build a new technique /algorithm that could help on improving the analysis for big data in social network
- Using any simulator / analyzing tools for testing and validating the proposed techniques /algorithms.

I will also explain my studies through the work flow plan. Therefore, my studies consist of multiple stages, each of which is related to the next stage which below :**Data Collection Stage**:

In this stage, massive amount of data can be collected from social network . It can be accessed individually or in combination or managed as a whole entity. These collected data is called data set.

Per-processing & features selection Stage:

In this stage, the data need to be ready for processing, so, some algorithms and mechanisms have to be applied on the collected data. Moreover, feature selection techniques are also necessary in this stage.

Classification and Clustering Stage:

Certainly, the data can be divided into several types such as text, image, audio, and video. These types have to processed separately. Thus, two techniques have to be used i.e., classification and clustering.

Data Mining (extraction) Stage:

In this stage, data mining process has to be used in order to find useful information that can be easily understood in large datasets.

Data Analyzing Stage:

Here, different analyses and logical reasoning have to be used so as to examine each component of the provided data from the previous stages to get a correct data for daily life.

Evaluation Stage:

This stage is concerned to test the validity of the proposed framework including the developed algorithms and check its ability to be applied in real applications. In addition in this stage, comparing the results of the proposed work to some recent related works has to be done in order to check the superiority and efficiency of the proposed work.

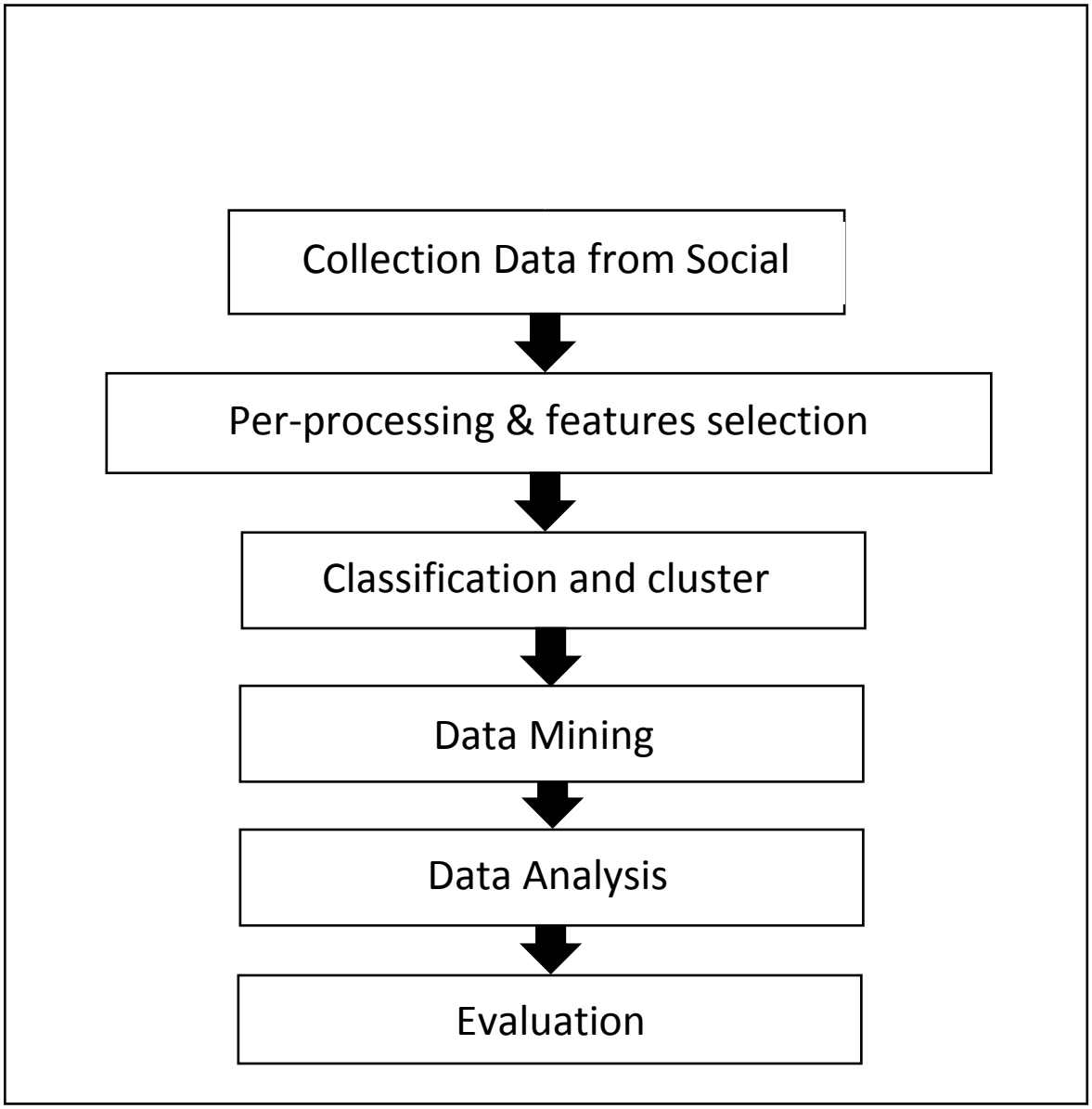

Fig (1): flow diagram of the workflow

Significance of the study

In fact, this field of study is very essential today as a result analysis for big data usage widespread in many scientific and economic trends. Therefore, this study is to investigate and trace the scientific achievement in analysis for big data and the ways of inventing effective techniques in order to analyze big data. It is, also, to get us familiar with analysis for big data and challenges and how to do analysis for big data.

Briefly, this study is expected to support the area of big data in order to build and provide an analysis of big data.

Proposed work plan

Year –wise plan of work and targets to be achieved

Year 1:

- Theoretical analysis of the existing big data analysis techniques.

Year 2:

- Designing and implementing a new model for big data analysis.
- Setting up the experimental results and performing analysis.

Year 3:

- Assessment of experimental setup in terms of comparison.
- Preparation of final synopsis and thesis.
- Submission of synopsis and thesis.

References

[1] Aggarwal, Charu C. "An introduction to social network data analytics." Social network data analytics. Springer, Boston, MA, 2011. 1-15.

[2] Bobade, Varsha B. "Survey paper on big data and Hadoop." *Int. Res. J. Eng. Technol.(IRJET)* 3.01 (2016)..

[3] Taouli, Amina&Djamel, amarbensaber&Keskes, Nabil &Bencherif, Khayra& Hassan, Badir. (2018). Semantic for Big Data Analysis: A survey.

[4] Nizam, Tasleem, and Syed Imtiyaz Hassan. "Big Data: A Survey Paper on Big Data Innovation and its Technology." International Journal of Advanced Research in Computer Science 8.5 (2017).

[5] Sutinen, Erkki, and JormaTarhio. "Filtration with q-samples in approximate string matching." Annual Symposium on Combinatorial Pattern Matching. Springer, Berlin, Heidelberg, 1996.

[6] Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." Ain Shams engineering journal 5.4 (2014): 1093-1113.

[7] Ahmed, Khaled, Neamat El Tazi, and Ahmad Hany Hossny. "Sentiment analysis over social Networks: An overview." 2015 IEEE International Conference on Systems, Man, and Cybernetics. IEEE, 2015.

[8] Perikos, Isidoros, and IoannisHatzilygeroudis. "A framework for analyzing big social data and modelling emotions in social media." 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService). IEEE, 2018.

[9] Hu, Qingwu, and Yuan Zhang. "An effective selecting approach for social media big data analysis—Taking commercial hotspot exploration with Weibo check-in data as an example." 2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA). IEEE, 2018.

[10] Vatrapu, Ravi, et al. "Social set analysis: A set theoretical approach to big data analytics." Ieee Access 4 (2016): 2542-2571.

[11] Huber, Marco F., Martin Voigt, and Axel-CyrilleNgongaNgomo. "Big data architecture for the semantic analysis of complex events in manufacturing." Informatik 2016 (2016).

[12] Chhabra, Tamanna, and JormaTarhio. "A filtration method for order-preserving matching." Information Processing Letters116.2 (2016): 71-74.

[13] Lu, Jiangbo, et al. "PatchMatch filter: edge-aware filtering meets randomized search for visual correspondence." IEEE transactions on pattern analysis and machine intelligence 39.9 (2016): 1866-1879.

[14] Nizam, Tasleem, and Syed Imtiyaz Hassan. "Big Data: A Survey Paper on Big Data Innovation and its Technology." International Journal of Advanced Research in Computer Science 8.5 (2017).

[15] Markić, Brano, Sanja Bijakšić, and Arnela Bevanda. "SENTIMENT ANALYSIS OF SOCIAL NETWORKS AS A CHALLENGE TO THE DIGITAL MARKETING." Ekonomski vjesnik/Econviews-Review of Contemporary Business, Entrepreneurship and Economic Issues 29.1 (2016): 95-107.

[16] Maddumage, Chamila, and M. P. Dhanushika. "R programming for Social Network Analysis-A Review." 2018 3rd International Conference on Information Technology

[17] Hu, Jie, Mengchi Liu, and Junchi Zhang. "A semantic model for academic social network analysis." Proceedings of the 2014 EEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE Press, 2014.

[18] Leenen, Louise, and Thomas Meyer. "Semantic technologies and big data analytics for cyber defence." Information Retrieval and Management: Concepts, Methodologies, Tools, and Applications. IGI Global, 2018. 1375-1388.

Research scholar

Mohammed Ali Mohammed Ali

Research Guide

DR . S.N.LOKHANDE
Assistant Professor
School of Computational Science,
SRTMU, Nanded